# Chapter 1

*Frederick Boehm*

*March 10, 2019*

## Todo list



## List of Figures



## List of Tables

Quantitative trait locus (QTL) studies in model organisms like mice can identify genomic regions that affect quantitative traits, such as systolic blood pressure and body weight. Tracing its origins back to Sax (1923), a genome-wide QTL "scan" discovers associations between genotypes and phenotypes by considering every position, one at a time, as a candidate QTL for the trait of interest. A region with strong evidence of association, then, defines a QTL. Because nearby markers have correlated genotypes, a QTL in a two-parent

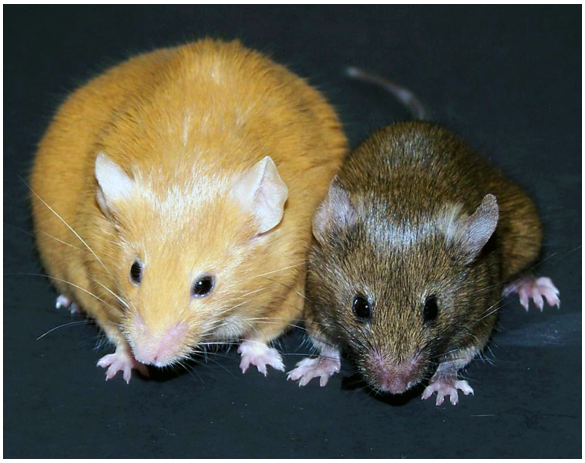

Figure 1: https://upload.wikimedia.org/wikipedia/commons/4/4d/Agouti_Mice.jpg

cross often spans multiple megabases in length and may contain more than a hundred genes. Identification of the causal gene from among those genes contained in the QTL is challenging and may require costly and time-consuming experiments. Growing needs for greater QTL mapping resolution fueled development, over the last two decades, of model organism multiparental populations for high-resolution QTL mapping.

Testing pleiotropy vs. separate QTL in model organisms can aid understanding of genetic architecture and reveal insights into biomolecular interactions in diverse areas including metagenomics, metabolomics, and behavioral genetics. The need to test pleiotropy arises when two (or more) traits demonstrate evidence of univariate QTL in a single, shared genomic region.

Identifying a gene that affects multiple traits may inform scientific understanding of interactions between biomolecules and ultimately contribute insights that aid development of new therapeutics. For example, mouse studies identified multiple biological roles for products of the *Agouti* gene. Mutations in the *Agouti* gene may lead to both yellow hair (in mice that are typically black) and obesity (Attie, Churchill, and Nadeau, 2017). Subsequent investigations uncovered two related biological roles for the Agouti protein. It antagonizes the action of $\alpha$-melanocyte-stimulating hormone both to prevent melanocyte-based melanin production and to disrupt melanocortin-4 receptor signaling in the brain. The former leads to yellow hair, while the latter causes weight gain. Later research identified altered signaling by the melanocortin-4 receptor in the brain as a leading cause of inherited obesity in humans. Therapeutics to mitigate the effects of *Agouti* mutations are currently being studied.

DNA variants that affect a trait like systolic blood pressure can be identified by systematically looking for correlations with genotypes at different positions across the genome.

Complex traits include clinical measurements, such as systolic blood pressure and body weight, as well as

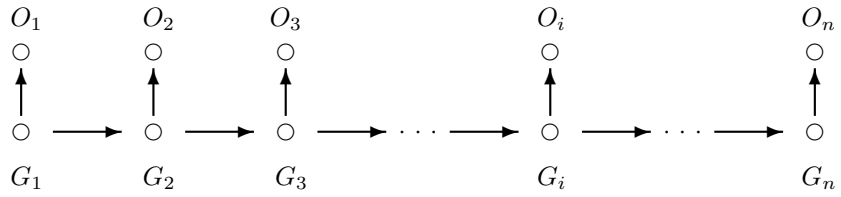


Figure 2: Illustration of a hidden Markov model. $G$'s indicate underlying genotypes; $O$'s indicate observed marker phenotypes.

newly measurable biomolecular traits like gene expression levels, protein concentrations, and lipid levels. Understanding the genetics of complex traits may inform the fields of biology, medicine, and public health.

Multiparental populations of model organisms are designed with the goals of incorporating many genetic variants and enhancing QTL mapping resolution. For much of the 20th century, researchers used two-parent crosses to identify large QTL that affect complex traits. Beginning in the late 1900s and early 2000s, geneticists recognized the limitations of two-parent QTL study designs. Due to the limited number of cumulative recombination events, existing two-parent study designs provided limited mapping resolution. Additionally, crosses of two inbred lines limits the collection of genetic variants in the offspring. A study design that incorporates DNA from more than two inbred lines would capture a greater number of genetic variants. Since mapping resolution is related to the cumulative number of meiotic crossover events on a chromosome, it would also be advantageous to have multi-generational mapping populations, in contrast to the traditional two-generation populations that result from backcrosses and intercrosses in two-parent designs.

Below, I present background material on statistical methods in quantitative trait locus (QTL) mapping. AfterI begin with an overview of QTL mapping in two-parent crosses. I then discuss pleiotropy testing in two-parent crosses before considering newly developed multiparental populations. I present the design for one multiparental population, the Diversity Outbred mice. I close the first chapter by arguing for the need for a pleiotropy test for multiparental populations.

## Two-parent crosses

## QTL mapping in two-parent crosses

QTL mapping is a systematic, statistical approach to identifying genetic loci where genetic variation affects phenotypic variation in a measured trait. We use the term "locus" to refer to a small, contiguous genomic region, often several megabases in length. Standard inputs are genome-wide marker genotypes for a collection of study subjects and a set of trait measurements on the same subjects.

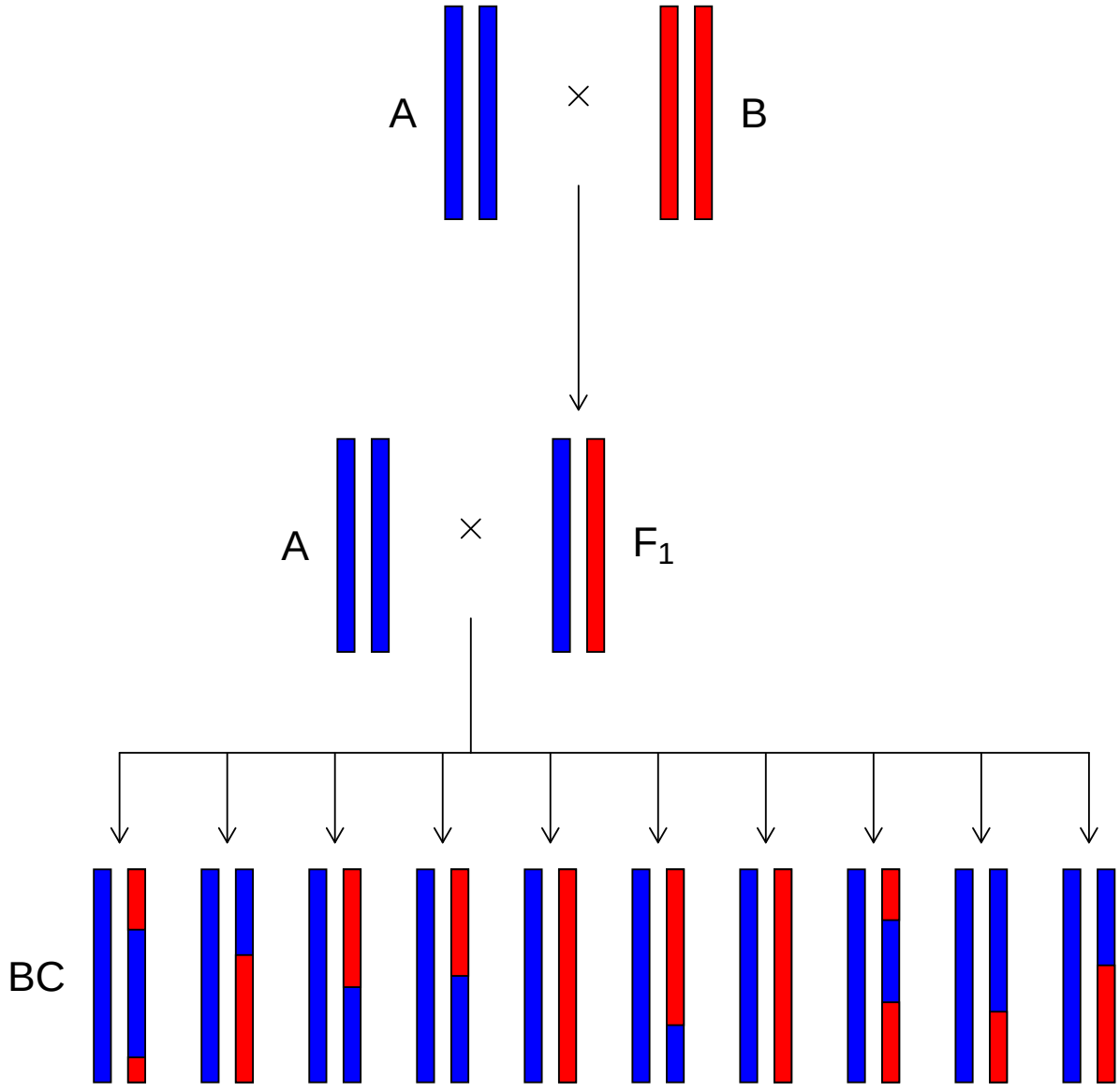


Figure 3: Breeding scheme for a backcross. Each pair of autosomes represents a single subject.

A univariate QTL scan is a procedure to interrogate the entire genome for genetic variants that affect a specified trait of interest. It requires as inputs genome-wide marker genotypes (or, equivalently, genotype probabilities) and trait measurements for a collection of subjects. One specifies a statistical model and calculates the likelihood of the model parameters given the observed (trait and marker) data at every marker. In studies with dense marker sets, it may suffice to calculate likelihoods only at the markers. However, for studies with fewer markers, investigators often benefit from inferring genotype probabilities at inter-marker positions.

After obtaining likelihoods for all markers, likelihood ratio test statistics are calculated at every marker. The inputs for these calculations are the likelihoods from the model fits at every marker and the likelihood for the null model, which contains no genotype data. The resulting likelihood ratios compare, for every marker, the null hypothesis that there is no QTL (at that marker) against the alternative that there is a QTL (at the specified marker). In other words, one performs a likelihood ratio test for every marker across the genome. Often this amounts to thousands of (statistically dependent) hypothesis tests. One then uses a permutation test to determine a genome-wide critical value for the likelihood ratio test statistic (Churchill and Doerge, 1994). Those loci for which the likelihood ratio test statistic is sufficiently large are declared QTL.

Modern QTL studies have their origins in the work of Sax (1923). Sax (1923), working 30 years before Watson and Crick (1953) reported the molecular structure of DNA, studied seed color and weight in the common bean (*Phaseolus vulgaris*). He used seed color to partition beans into genetic marker classes and identified an association between marker class and bean weight. Implicitly, Sax (1923) leveraged the association between marker genotype class and seed color to assign beans to genetic marker classes without having explicit genotype data.

Lander and Botstein (1989), in a landmark methodology report, presented interval mapping as a strategy to identify QTL. In interval mapping, they used the correlations between genetic markers to infer genotype probabilities between markers. They then use an expectation maximization algorithm (Dempster, Laird, and Rubin, 1977; Lander and Green, 1987) to fit a statistical mixture model with a normally distributed component for each of three genotype classes. They summarized the evidence for a QTL at each position with the log odds (LOD) score statistic. The LOD score is the base 10 log likelihood ratio test statistic for the competing hypotheses of the presence of a QTL at the candidate position against the null hypothesis of no QTL at the candidate position. The work of Lander and Botstein (1989) fueled interest in QTL mapping from both biologists and statisticians. Many of the open questions that Lander and Botstein (1989) present, including approaches for defining QTL endpoints, accounting for multiple hypothesis tests, and allowing for genotyping errors when inferring genotype probabilities, are active areas of research 30 years after their

publication.

Two major statistical challenges are what Karl W Broman and Sen (2009) call the "missing data" problem and the "model selection" problem. The "missing data" problem arises in QTL studies because genotypes are obtained at only select markers. In this sense, genotypes at positions between markers are "missing" because they aren't explicitly measured. A flexible framework for resolving the "missing data" problem involves a hidden Markov model (Karl W Broman and Sen, 2009; Karl W. Broman, 2006). In this hidden Markov model, marker genotypes are observed random variables, while genotypes at intervening bases are unobserved random variables (Karl W Broman and Sen, 2009; Karl W. Broman, 2006). Two sets of recursive equations, termed "forward" and "backward" equations, enable efficient calculation of genotype probabilities (conditional on the observed marker genotypes) (Baum et al., 1970).

A QTL analysis involves choosing a statistical model to be fitted at every marker. As in any statistical modeling, there is not one best model selection procedure. One needs to decide which main effects and interactions are needed. Also, one often transforms the trait values to achieve approximate normality before performing the QTL scan.

Many investigators consider multiple models before performing a QTL scan. For simplicity, one may choose a linear model that contains additive effects for the minor allele count (0, 1 or 2) at a given marker (Equation 1) (Martinez and Curnow, 1992; Haley and Knott, 1992).

$$\text{trait} = \text{mean for major allele homozygotes} + (\text{minor allele count})(\text{minor allele effect}) + \text{random error} \quad (1)$$

Assuming that the random errors are normally distributed (and independent with common variance $\sigma^2$), one may use the statistical technique called "ordinary least squares" to fit the model and to solve for $\hat{b}$ and $\hat{\sigma}^2$. In "ordinary least squares", one solves for the set of parameter values that minimize the residual sum of squares. In the above equation, our two parameters are the minor allele effect and the random error variance. The residual sum of squares is an expression that tells us how far each observed data point is from its predicted value for a set of specified parameter values. Equation **??** defines residual sum of squares for a univariate QTL analysis.

$$\text{residual sum of squares} = (\text{fitted trait value} - \text{observed trait value})^2 \quad (2)$$

In anticipation that multivariate mapping of correlated traits would enhance statistical power to detect QTL and would improve precision of QTL positions, both Jiang and Zeng (1995) and Korol, Ronin, and Kirzhner (1995) developed multivariate interval mapping procedures.

say more here

In the context of multivariate QTL mapping, Jiang and Zeng (1995) proposed and developed a statistical hypothesis test for distinguishing a single pleiotropic locus that affects multiple traits from two (or more) separate QTL. Their test applies to the setting in which multiple traits map to a single genomic region (Jiang and Zeng, 1995). That region, then, can be interrogated further by asking whether it harbors a single pleiotropic QTL or multiple separate QTL. For simplicity, we focus on the case where two traits both map to a single genomic region, and we wish to determine whether the two traits share a single QTL or have two separate QTL.

## Multiparental populations and Diversity Outbred mice

Near the turn of the century, geneticists sought a mammalian gene mapping resource that could be used for study of a wide variety of quantitative traits. The magnitude of such an undertaking required a collaborative, community-supported approach (Koning and McIntyre, 2014). Scientists conceived of the Diversity Outbred (DO) mouse population as such a high-resolution gene mapping resource. They elected to seed the population with partially inbred progenitors of the Collaborative Cross mouse population. The Collaborative Cross mating design started with mice from eight inbred lines: A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ, NZO/HILtJ, CAST/EiJ, PWK/PhJ, WSB/EiJ.

The designers of the Collaborative Cross used a "multi-funnel" mating scheme to generate mice with DNA from all eight founder lines over the course of 3 generations. For example, in one funnel, mating pairs are: A x B, C x D, E x F, and G x H in the first generation. AB offspring would then mate with CD offspring and EF offspring would mate with GH mice. Finally, the ABCD mice would mate with the EFGH mice to create a generation of mice that contain genetic material from all eight inbred founder lines. Subsequent generations of inbreeding resulted in multiple inbred lines for the Collaborative Cross.

make this figure

The designers of the Diversity Outbred population started with

get number of mice and from which generations

Due to the breeding scheme, each Diversity Outbred mouse is a highly heterozygous mosaic of founders' DNA

(Figure). With each generation, the Diversity Outbred mice accumulate meiotic recombinations. This is because each mouse inherits DNA from its parents, and each meiosis provides opportunities for recombinations.

add figure with mosaics over multiple generations here

need to explain this better. How does Andrew explain it in his thesis?

The cumulative effect of recombinations over many generations of outbreeding is to create mice whose DNA mosaic contains smaller and smaller contiguous pieces from each founder line.

The Diversity Outbred mice enable high-resolution QTL mapping (Gatti et al., 2014). One factor that contributes to high-resolution mapping is the ability to infer the founder line from which each marker's DNA arose.

## QTL mapping in DO mice

QTL mapping in Diversity Outbred mice, like that in mice from two-parent crosses, is a multi-step procedure: 1. data acquisition, 2. inference of missing genotypes, and 3. modeling phenotypes as a function of genotypes. Data acquisition involves measurement of phenotypes and, at specified genetic markers, termed "single nucleotide polymorphism" or SNP markers, measurement of two-allele genotypes. Often, the SNP marker genotypes are obtained by use of a microarray, such as the GigaMUGA SNP microarray (Morgan et al., 2015).

The next step, missing genotypes inference, is needed because of the "missing data problem" (Karl W Broman and Sen, 2009). It takes as input the two-allele genotypes at the measured SNP markers. An expectation maximization algorithm (Dempster, Laird, and Rubin, 1977) for a hidden Markov model, developed by Karl W Broman (2012b) and Karl W Broman (2012a) and implemented in the `qtl2` R package (K. W. Broman et al., 2019), outputs 36-state genotype probabilities for all nuclear autosomal markers and pseudomarkers. Pseudomarkers, as we use the term, are arbitrary nucleotide bases at which the researcher wants 36-state genotype probabilities. Finally, we collapse the 36-state genotype probabilities to eight founder allele dosages at each marker. This last step is optional, but often helpful, because the simplified models require specification of fewer parameters. We then treat the founder allele dosages as known quantities in subsequent steps.

After inferring founder allele dosages, we address the second major statistical challenge of QTL mapping: the "model selection" problem (Karl W Broman and Sen, 2009). Linear regression is widely used because of its ease of implementation, computational speed, and interpretability of results. A genetically "additive" linear model, in which we assume a linear relationship between a trait and each founder allele's dosage, is widely used (Gatti et al., 2014; K. W. Broman et al., 2019). The R package `qtl2` provides a straightforward user

interface to include covariates and interaction terms in the linear model (K. W. Broman et al., 2019).

## Testing pleiotropy in two-parent crosses

Jiang and Zeng (1995) developed a suite of methods for analyzing multivariate phenotypes in two-parent crosses. Among the novel methods in their article is a test of pleiotropy vs separate QTL. They explain that such a test is useful when two traits map to a single genomic region. The question then arises "do the two traits associate with the same locus, or do they associate with separate loci"?

If both traits associate with the same locus, then that locus is called "pleiotropic". If the two traits associate with separate loci, then we say that there are separate QTL (one for each trait). While "pleiotropy" has multiple related usages in the genetics literature, we use "pleiotropy" here to refer to the situation in which one genetic locus affects two or more traits.

In the development of their test of pleiotropy v separate QTL, Jiang and Zeng (1995) model the multivariate phenotype as the sum of a linear function of the genotypes and a random error term. They assume that the phenotypes matrix is related to the genotypes data (for a single marker) through the following equation:

$$Y = XB + E \tag{3}$$

Where Y is a n by 2 matrix of phenotype values (with each row being one subject and each column being one trait) , X is a n by 1 matrix of genotypes and B is a 1 by 2 matrix of allele effects. The random error is assumed to follow a normal distribution with mean zero and a positive variance. Jiang and Zeng (1995) also incorporate additional terms to model dominant effects and to control for residual genetic variation. We forego inclusion of dominant terms and additional markers for controlling residual genetic variation, although, in our linear mixed effects model, we do include polygenic random effects. Extensions of our model and software to accommodate dominance effects are straightforward.

Jiang and Zeng (1995) use a mixture model to relate genotypes to phenotypes. They observe that the there are 9 possible ordered pairs of genotypes for two markers in a F2 population, since each marker has one of three genotypes (AA, AB, BB). They thus approach the problem as one with a 9-state mixture model. They provide the equations needed for an expectation-conditional maximization (ECM) algorithm for fitting the mixture model. They use a chi-square distribution with one degree of freedom as the null distribution of the test statistic.

Further investigations by Jiang and Zeng (1995) with simulations revealed reasonable behavior of the hypothesis test under the specified conditions. They learned that dense marker coverage tends to aid in distinguishing pleiotropy from separate QTL.

IS THIS TRUE?? look back at JZ 1995's simulations. How did they examine power and type I error rates? Do they examine univariate LOD strength when considering power? What about interlocus distance?

They use a computational algorithm called "expectation - conditional maximization" to find these values. ECM is an iterative algorithm. The user inputs starting values for the parameters and the algorithm changes them one at a time until the change in log likelihood between successive iterations is sufficiently small. The algorithm is then said to have "converged". The final iteration's parameter values are treated as those that maximize the likelihood.

Jiang and Zeng (1995), in the context of their multivariate QTL mapping methods, describe a test of pleiotropy vs. separate QTL. They formulate a test in which the null hypothesis is pleiotropy while the the alternative hypothesis is separate QTL. In other words, they consider the case in which the two traits map to the same locus (pleiotropy) as a special case of the more general setting (separate QTL) in which the two traits may or may not map to the same locus.

In statistical terminology, one would say that the parameter space is restricted under the null hypothesis. Here, the parameter space is the collection of ordered pairs of markers in the genomic region of interest. The restriction of that collection under pleiotropy corresponds to limiting consideration to only those ordered pairs that have both traits mapping to a single locus.

```
## - Attaching packages --------------------- tidyverse 1.2.1 -

## v ggplot2 3.1.0      v purrr  0.3.1
## v tibble  2.0.1      v dplyr  0.8.0.1
## v tidyr  0.8.3      v stringr 1.4.0
## v readr  1.3.1      v forcats 0.4.0

## - Conflicts ---------------------- tidyverse_conflicts() -
## x .GlobalEnv::cross() masks purrr::cross()
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()        masks stats::lag()
```

Jiang and Zeng (1995) then calculate the likelihoods of the two models (that under the hypothesis of pleiotropy

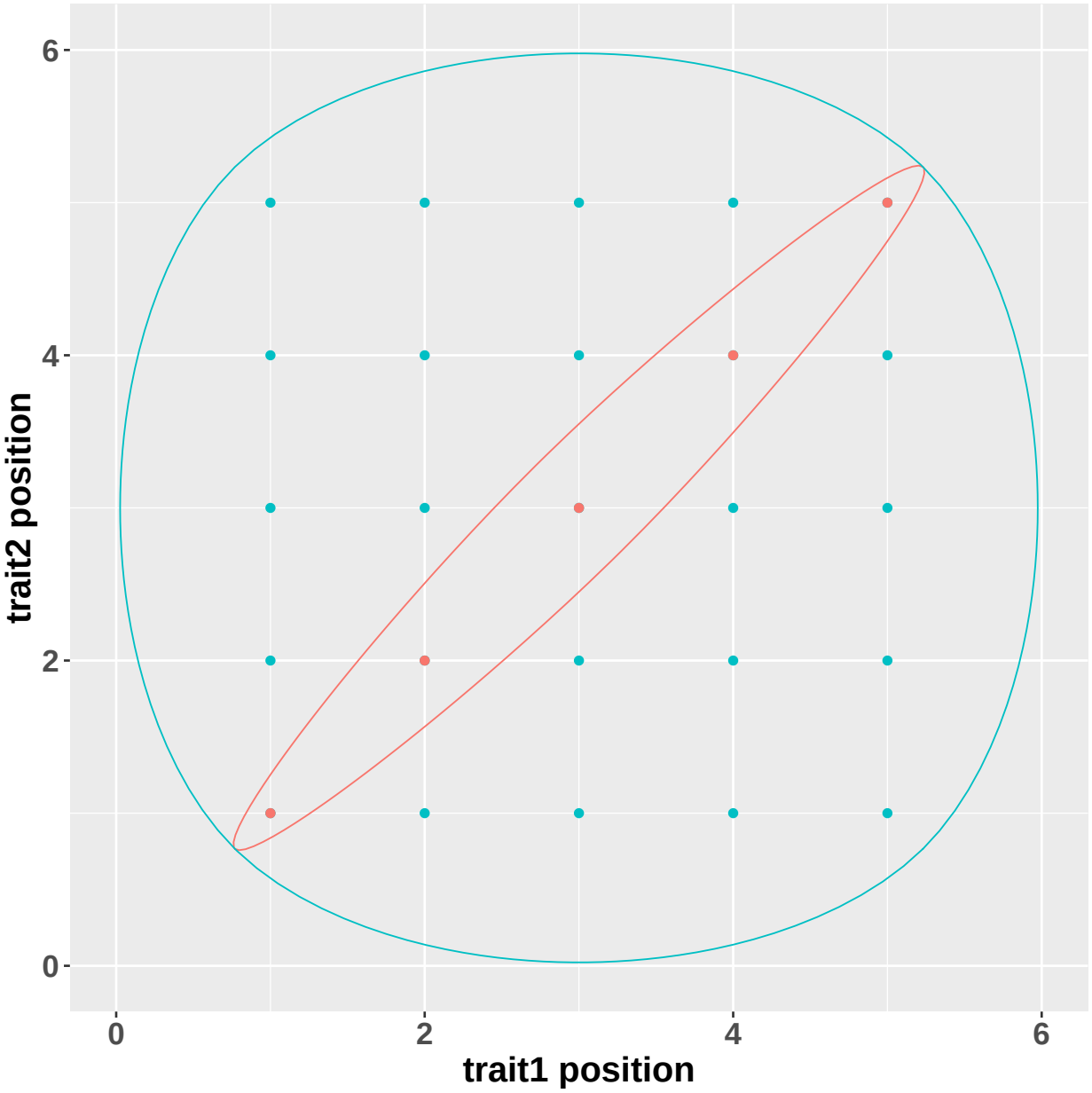


Figure 4: Two-dimensional grid of ordered pairs of markers.

and that under the hypothesis of separate QTL) at each ordered pair of putative loci in the genomic region of interest. The likelihood ratio test statistic is the logarithm of the ratio of the maximum of the likelihoods under pleiotropy to the maximum of the likelihoods under the separate QTL hypothesis.

Jiang and Zeng (1995) determined p-values for their test statistics by comparing them to a chi-squared distribution with 1 degree of freedom.

## References


Attie, Alan D, Gary A Churchill, and Joseph H Nadeau (2017). "How mice are indispensable for understanding obesity and diabetes genetics". In: *Current opinion in endocrinology, diabetes, and obesity* 24.2, p. 83.

Baum, Leonard E, Ted Petrie, George Soules, and Norman Weiss (1970). "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains". In: *The annals of mathematical statistics* 41.1, pp. 164–171.

Broman, K. W., D. M. Gatti, P. Simecek, N. A. Furlotte, P. Prins, S. Sen, B. S. Yandell, and G. A. Churchill (2019). "R/qtl2: Software for mapping quantitative trait loci with high-dimensional data and multi-parent populations". In: *Genetics, to appear*.

Broman, Karl W (2012a). "Genotype probabilities at intermediate generations in the construction of recombinant inbred lines". In: *Genetics* 190.2, pp. 403–412.

– (2012b). "Haplotype probabilities in advanced intercross populations". In: *G3: Genes, Genomes, Genetics* 2.2, pp. 199–202.

Broman, Karl W. (2006). *Use of hidden Markov models for QTL mapping*. Tech. rep. 125. Johns Hopkins University Department of Biostatistics. URL: http://biostats.bepress.com/jhubiostat/paper125.

Broman, Karl W and Saunak Sen (2009). *A Guide to QTL Mapping with R/qtl*. Vol. 46. Springer.

Churchill, Gary A and Rebecca W Doerge (1994). "Empirical threshold values for quantitative trait mapping." In: *Genetics* 138.3, pp. 963–971.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38.

Gatti, Daniel M, Karen L Svenson, Andrey Shabalin, Long-Yang Wu, William Valdar, Petr Simecek, Neal Goodwin, Riyan Cheng, Daniel Pomp, Abraham Palmer, et al. (2014). "Quantitative trait locus mapping methods for diversity outbred mice". In: *G3: Genes, Genomes, Genetics* 4.9, pp. 1623–1633.

Haley, Chris S and Sarah A Knott (1992). "A simple regression method for mapping quantitative trait loci in line crosses using flanking markers". In: *Heredity* 69.4, pp. 315–324.

Jiang, Changjian and Zhao-Bang Zeng (1995). "Multiple trait analysis of genetic mapping for quantitative trait loci." In: *Genetics* 140.3, pp. 1111–1127.

Koning, Dirk-Jan de and Lauren M McIntyre (2014). "GENETICS and G3: Community-Driven Science, Community-Driven Journals". In: *Genetics* 198.1, pp. 1–2.

Korol, Abraham B, Yefim I Ronin, and Valery M Kirzhner (1995). "Interval mapping of quantitative trait loci employing correlated trait complexes." In: *Genetics* 140.3, pp. 1137–1147.

Lander, Eric S and David Botstein (1989). "Mapping mendelian factors underlying quantitative traits using RFLP linkage maps." In: *Genetics* 121.1, pp. 185–199.

Lander, Eric S and Philip Green (1987). "Construction of multilocus genetic linkage maps in humans". In: *Proceedings of the National Academy of Sciences* 84.8, pp. 2363–2367.

Martinez, O and RN Curnow (1992). "Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers". In: *Theoretical and Applied Genetics* 85.4, pp. 480–488.

Morgan, Andrew P, Chen-Ping Fu, Chia-Yu Kao, Catherine E Welsh, John P Didion, Liran Yadgary, Leeanna Hyacinth, Martin T Ferris, Timothy A Bell, Darla R Miller, et al. (2015). "The mouse universal genotyping array: from substrains to subspecies". In: *G3: Genes| Genomes| Genetics*, g3–115.

Sax, Karl (1923). "The association of size differences with seed-coat pattern and pigmentation in Phaseolus vulgaris". In: *Genetics* 8.6, p. 552.

Watson, James D and Francis HC Crick (1953). "Molecular structure of nucleic acids". In: *Nature* 171.4356, pp. 737–738.